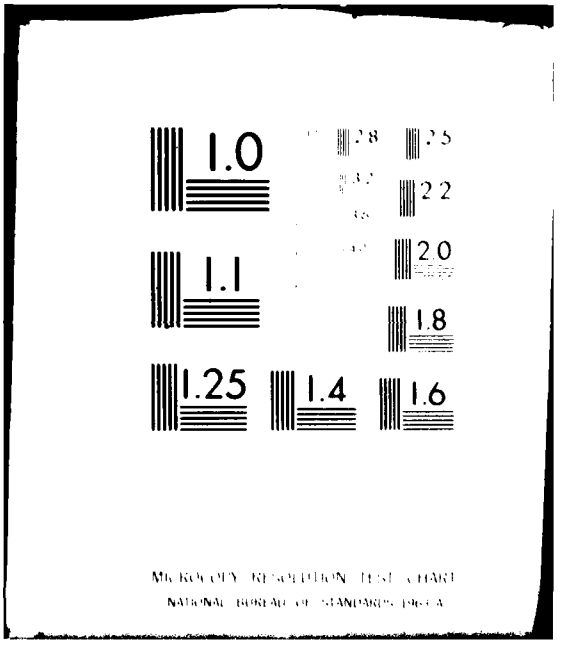
AD-A088 305 CORNELL UNIV ITHACA N Y SCHOOL OF OPERATIONS RESEARC--ETC F/G 12/1
A NOTE ON THE LOWER BOUND FOR THE P(CS) OF GUPTA'S SUBSET SELEC--ETC(U)
DEC 79 R E BECHHOFER, T J SANTNER DAAG29-80-C-0036
UNCLASSIFIED TR-401 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

LEVEL II

13

SCHOOL OF OPERATIONS RESEARCH AND INDUSTRIAL ENGINEERING

COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK

TECHNICAL REPORT NO. 401

January 1979
(Revised December 1979)

A NOTE ON THE LOWER BOUND FOR THE P(CS) OF GUPTA'S SUBSET SELECTION PROCEDURE

by

Robert E. Bechhofer and Thomas J. Santner

A

Research supported by

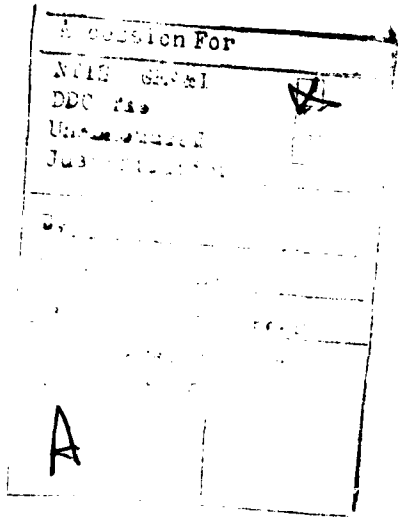
U.S. Army Research Office - Durham contract DAAG29-80-C-0036,

Office of Naval Research Contract N00014-75-C-0586,

and

National Science Foundation grant ENG-7906914

at Cornell University



A NOTE ON THE LOWER BOUND FOR THE P{CS} OF GUPTA'S SUBSET SELECTION PROCEDURE

Robert E. Bechhofer and Thomas J. Santner

Cornell University, Ithaca, New York 14850

ABSTRACT

The lower bound on the specified P^* for the Gupta procedure for selecting a subset containing the best of k populations, and for the Gupta-Sobel procedure for selecting a subset containing all populations at least as good as a control population is studied via impartial "no data" minimax decision rules. Gibbons, Olkin, and Sobel (1977) state that a theoretical lower bound is $1/2^k$ for both problems. Our analysis shows (a) that $1/k$ is the correct lower bound for the first problem, and (b) that $1/2^k$ is the correct lower bound for the second problem provided that a particular loss function is adopted. Other (reasonable) choices of loss function lead to different lower bounds for the second problem.

1. SUBSET SELECTION OF THE BEST POPULATION

Gupta's subset selection procedure (Gupta, 1956, 1965) selects a non-empty (small) subset from $k \geq 2$ populations such that the probability is at least equal to a specified value P^*

that the "best" population is contained in the selected subset. His 1965 paper mentions (p. 230) that "...for values of $P^* \leq 1/k$ there always exists a no data decision rule." In their recent text Gibbons, Olkin and Sobel (1977) (GOS) state that the theoretical lower bound for P^* is $1/2^k$. This note gives arguments showing that Gupta's value of $1/k$ is the correct lower bound.

For simplicity, the computations that follow assume that if θ_i is the parameter characterizing the "goodness" of the ith population $(1 \leq i \leq k)$, then the values of the elements of $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Omega$ are distinct so that there is a unique best population. A correct selection is made only when the selected subset contains the best population.

In the "no data" situation a selection rule is described in terms of a sampling scheme $\{p_{ij} | 1 \leq i \leq k, 1 \leq j \leq \binom{k}{i}\}$ for choosing a non-empty subset. Here p_{ij} denotes the probability of selecting the jth subset of size i where the $\binom{k}{i}$ subsets are written in some fixed order. We assume that the statistician will use an impartial (i.e., invariant wrt the group of permutations) rule (Bahadur and Goodman, 1952 and Eaton, 1967); it is easy to check that for the case of zero-one loss this is equivalent to requiring that $p_{ij_1} = p_{ij_2}$ for $1 \leq i \leq k$ and $1 \leq j_1, j_2 \leq \binom{k}{i} \iff p_{ij} = p_i / \binom{k}{i}$ for $1 \leq i \leq k$, $1 \leq j \leq \binom{k}{i}$ where $\underset{\sim}{p} = (p_1, p_2, \ldots, p_k)$ satisfies $p_i \geq 0$ $(1 \leq i \leq k)$ and $\sum_{i=1}^{k} p_i = 1$. Hence, impartial decision rules operate in two independent stages: First a subset of size i is chosen according to $\underset{\sim}{p}$, and then one of the $\binom{k}{i}$ subsets of size i is chosen at random. Both the expected subset size, $E_{\underset{\sim}{\theta}}\{S|\underset{\sim}{p}\}$, and the probability of a correct selection, $P_{\underset{\sim}{\theta}}\{CS|\underset{\sim}{p}\}$, are independent of $\underset{\sim}{\theta}$ for invariant no-data rules $\underset{\sim}{p}$. We obtain $P\{CS|\underset{\sim}{p}\} \equiv \inf_{\Omega} P_{\underset{\sim}{\theta}}\{CS|\underset{\sim}{p}\} = \sum_{i=1}^{k} p_i(i/k)$, and $E\{S|\underset{\sim}{p}\} \equiv \sup_{\Omega} E_{\underset{\sim}{\theta}}\{S|\underset{\sim}{p}\} = \sum_{i=1}^{k} i p_i = kP\{CS|\underset{\sim}{p}\}$. Any choice of $P^* \in [1/k, 1]$ can be attained exactly by an appropriate rule. Table I lists the values of $P\{CS|\underset{\sim}{p}\}$ and $E\{S|\underset{\sim}{p}\}$

TABLE I

Some Possible Choices of $\underset{\sim}{p}$, and Associated Performance Characteristics

$\underset{\sim}{p}$	$\inf_{\Omega} P_{\underset{\sim}{\theta}}\{CS\|\underset{\sim}{p}\}$	$\sup_{\Omega} E_{\underset{\sim}{\theta}}\{S\|\underset{\sim}{p}\}$
$p_k = 1$	1	k
$p_i = 1/k \quad (1 \le i \le k)$	$(k+1)/2k$	$(k+1)/2$
$p_i = \binom{k}{i}/(2^k-1) \quad (1 \le i \le k)$	$2^{k-1}/(2^k-1)$	$k2^{k-1}/(2^k-1)$
$p_1 = 1$	$1/k$	1

for several $\underset{\sim}{p}$.

Suppose that $E_{\underset{\sim}{\theta}}\{S|\delta\}$ is regarded as the risk associated with an arbitrary rule δ. The rule $p_1 = 1$ is clearly the uniformly minimum risk procedure in the class of no-data rules; the value of $P\{CS|\underset{\sim}{p}\}$ associated with $p_1 = 1$ is $1/k$. In data problems the P^* condition, $\inf_{\Omega} P_{\underset{\sim}{\theta}}\{CS|\delta\} \ge P^*$, specifies a class of rules, $C(P^*)$, to be studied. The rule $p_1 = 1$ is in $C(P^*)$ for all $P^* \le 1/k$ and hence it is a uniformly minimum risk rule. P^* must be chosen greater than $1/k$ for the problem to require a data-dependent solution under the risk $E_{\underset{\sim}{\theta}}\{S|\delta\}$.

Remark 1.1. One could argue that the risk $E_{\underset{\sim}{\theta}}\{S|\delta\}$ is deficient and that it would be better to drop the P^* condition and adopt a loss function which takes into account both S and the event of correct selection. Goel and Rubin (1977) and Chernoff and Yahav (1977) give some examples of this latter approach. Our basic analysis is closer in spirit to that of classical statistical theory in which an optimal procedure is selected from a subclass of all procedures defined by some condition such as the size of a test or the unbiasedness of an estimator.

Remark 1.2. The statement on p. 297 of GOS to the effect that "...the probability $1/2^k$ can be attained by simply tossing a

fair coin for each population to determine whether to include it in the subset or not" is incorrect. This rule, which permits the choice of an empty subset (an option which is not allowed for this problem), yields $P_{\underset{\sim}{\theta}}\{CS\} = 1/2$.

Remark 1.3. GOS state (p. 297) that P^* should be chosen greater than $(1/2 + 1/2k)$ for applied work. Practitioners who follow the advice will automatically satisfy the correct theoretical bound since $1/2 + 1/2k > 1/k$ for $k \geq 2$.

2. SUBSET SELECTION WITH RESPECT TO A CONTROL

The Gupta-Sobel subset selection procedure (Gupta and Sobel, 1958) is used for selecting a subset (possibly empty) of k ($k \geq 1$) populations containing all those populations at least as good as a control. In the sequel below we call such populations "optimal." A correct selection is said to occur if and only if the above stated goal is achieved. GOS states (pp. 307,310) that the lower bound for P^* is $1/2^k$.

Our analysis of the present problem will differ from that of the previous one. The reason is that while S must clearly be minimized in Section 1, its role in the present problem is more complicated. If we write the parameter space Ω as $\bigcup_{j=0}^{k} \Omega_j$ where Ω_j $(0 \leq j \leq k)$ represents those points for which exactly j populations are at least as good as the control, then when $\underset{\sim}{\theta} \in \Omega_j$ we would hope that the selected subset contains all j optimal populations and no others (in which case $S = j$). In particular, when $\underset{\sim}{\theta} \in \Omega_0$ we would hope $S = 0$. Many loss functions can be proposed which embody the above notion; below we show that $1/2^k$ is the lower bound of the infimum of the probability of correct selection corresponding to a minimax rule for one of these loss functions. However, other reasonable loss functions yield different lower bounds.

We now introduce three possible such loss functions. For each $\underset{\sim}{\theta} \in \Omega$ let $B(\underset{\sim}{\theta}) = \{i | \theta_i \geq \theta_0\}$ denote the set of optimal

populations under $\underset{\sim}{\theta}$; also let R denote the set of selected populations. Define L_1, L_2, and L_3 as follows:

$$L_1(\underset{\sim}{\theta},R) = \begin{cases} 0 & \text{if } R = B(\underset{\sim}{\theta}) \\ 1 & \text{otherwise,} \end{cases}$$

$$L_2(\underset{\sim}{\theta},R) = \begin{cases} 0 & \text{if } R = B(\underset{\sim}{\theta}) \text{ and } \underset{\sim}{\theta} \in \Omega_0 \text{ or } R \supset B(\underset{\sim}{\theta}) \text{ and } \underset{\sim}{\theta} \in \Omega-\Omega_0 \\ 1 & \text{otherwise,} \end{cases}$$

$$L_3(\underset{\sim}{\theta},R) = |R-B(\underset{\sim}{\theta})|.$$

Here "-" denotes set difference and $|A|$ denotes the cardinality of the set A. L_1 penalizes the statistician unless S contains exactly the set of optimal populations. L_2 is a milder form of L_1 which dichotomizes the problem into $\underset{\sim}{\theta} \in \Omega_0$ and $\underset{\sim}{\theta} \in \Omega-\Omega_0$; for $\underset{\sim}{\theta} \in \Omega_0$ it penalizes the statistician unless the empty set is chosen while for $\underset{\sim}{\theta} \in \Omega-\Omega_0$ it penalizes him/her when the selected subset does not contain all optimal poulations (but S can contain non-optimal populations as well). Finally, L_3 is the number of populations in the selected subset which are worse than the control. Alternatively, one might even adopt a loss function which is, e.g., a linear combination of L_1 or L_2 and L_3.

We next introduce no data rules for our present problem; these are of the form $\underset{\sim}{p} = (p_0,p_1,\ldots,p_k)$ where a subset of size i $(0 \leq i \leq k)$ is chosen according to $\underset{\sim}{p}$, and then one of the $\binom{k}{i}$ subsets of size i is chosen at random. In contrast to the situation in Section 1, $P_{\underset{\sim}{\theta}}\{CS|\underset{\sim}{p}\}$ now depends on $\underset{\sim}{\theta} \in \Omega$ but is a constant over $\underset{\sim}{\theta} \in \Omega_j$ $(0 \leq j \leq k)$. A straightforward computation gives $P\{CS|\underset{\sim}{p}\} \equiv \inf_{\Omega} P_{\underset{\sim}{\theta}}\{CS|\underset{\sim}{p}\} = \min_{1\leq i\leq k} \sum_{j=1}^{k} p_j\binom{k-i}{j-i}/\binom{k}{j} = p_k$. For each loss L_i $(1 \leq i \leq 3)$, Table II lists (1) the maximum of the risk $R_i(\underset{\sim}{\theta},\underset{\sim}{p}) \equiv E_{\underset{\sim}{\theta}}\{L_i(\underset{\sim}{\theta},S)|\underset{\sim}{p}\}$ for an arbitrary no-data $\underset{\sim}{p}$; (2) the minimax rule, $\underset{\sim}{p}^{(i)}$; (3) the infimum of the probability of correct selection under $\underset{\sim}{p}^{(i)}$, $P\{CS|\underset{\sim}{p}^{(i)}\}$; and (4) the minimax risk.

It can be shown that the maximum risk under L_1 (L_2), $\sup_{\Omega} R_1(\underset{\sim}{\theta},\delta)$, $(\sup_{\Omega} R_2(\underset{\sim}{\theta},\delta))$ is at least $1 - 1/2^k$ $(1/2)$ for any data dependent rule, δ. Hence the statistician should not use a

TABLE II

Minimax Rules and Associated Risk for L_i $(i = 1,2,3)$

i	$\sup_{\Omega} R_i(\theta, \underset{\sim}{p})$	$\underset{\sim}{p}^{(i)}$	$P\{CS \mid \underset{\sim}{p}^{(i)}\}$	$\sup_{\Omega} R_i(\theta, \underset{\sim}{p}^{(i)})$
1	$1 - \min_{0 \le j \le k} \frac{p_j}{\binom{k}{j}}$	$p_j^{(1)} = \frac{\binom{k}{j}}{2^k}$ $(0 \le j \le k)$	$1/2^k$	$1 - 1/2^k$
2	$\max\{1-p_0, 1-p_k\}$	$p_0^{(2)} = 1/2 = p_k^{(2)}$	$1/2$	$1/2$
3	$\sum_{j=1}^{k} jp_j$	$p_0^{(3)} = 1$	0	0

data dependent rule if P^* is chosen less than or equal to $1/2^k$ $(1/2)$. Similarly the analysis of L_3 shows that P^* must be positive or else the no-data rule $\underset{\sim}{p}^{(3)}$ should be used.

ACKNOWLEDGMENT

This research was supported in part by U.S. Army Research Office-Durham contract DAAG29-80-C-0036, Office of Naval Research contract N00014-75-C-0586 and the National Science Foundation grant ENG-7906914 at Cornell University.

BIBLIOGRAPHY

Bahadur, R.R. and Goodman, L.A. (1952). Impartial decision rules and sufficient statistics. Annals of Math. Statist. 23, 553-62.

Chernoff, H. and Yahav, J. (1977). A subset selection problem employing a new criteria. Statistical Decision Theory and Related Topics II (S.S. Gupta and D.S. Moore, Eds.). New York: Academic Press, 93-120.

Eaton, M.L. (1967). The generalized variance, testing and ranking problem. Annals of Math. Statist. 38, 941-43.

Gibbons, J.D., Olkin, I. and Sobel, M. (1977). Selecting and Ordering Populations. Wiley-Interscience.

Goel, P.K. and Rubin, H. (1977). On selecting a subset containing the best population--a Bayesian approach. Annals of Statist. 5, 969-983.

Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Mimeo Series 150, Institute of Statistics, Univ. of North Carolina, Chapel Hill, N.C.

Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. Technometrics 7, 225-45.

Gupta, S.S. and Sobel, M. (1958). On selecting a subset which contains all populations better than a standard. Annals of Math. Statist. 29, 235-44.

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
#401	AD-A088 305	

4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED
A Note on the Lower Bound for the P{CS} of Gupta's Subset Selection Procedure	Technical Report
	6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
Robert E. Bechhofer and Thomas J. Santner	DAAG29-80-C-0036 N00014-75-C-0586 NSF ENG-7906914

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853	

11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE
National Science Foundation Washington, D.C. 20550	December 1979
	13. NUMBER OF PAGES 7

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Subset selection procedures, lower bound, selection with respect to a control

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The lower bound on the specified P^* for the Gupta procedure for selecting a subset containing the best of k populations, and for the Gupta-Sobel procedure for selecting a subset containing all populations at least as good as a control population is studied via impartial "no data" minimax decision rules. Gibbons, Olkin, and Sobel (1977) state that a theoretical lower bound is $1/2^{(k)}_{L}$ for both problems. Our analysis shows (a) that $1/k$ is the correct lower bound for the first problem, and (b) that $1/2^k$ is the

correct lower bound for the second problem provided that a particular loss function is adopted. Other (reasonable) choices of loss function lead to different lower bounds for the second problem.